Isotopic Labeling

DOI: 10.1002/ange.201303429

N-Terminal Labeling of Peptides by Trypsin-Catalyzed Ligation for Quantitative Proteomics**

Yanbo Pan, Mingliang Ye, Liang Zhao, Kai Cheng, Mingming Dong, Chunxia Song, Hongqiang Qin, Fangjun Wang, and Hanfa Zou**

Labeling tryptic peptides with stable isotopes is one of the most important methods for quantitative proteomics, and many ingenious chemical labeling strategies have been developed.[1] Incorporation of one isotopically labeled tag onto a peptide terminus represents an ideal labeling approach, as it would simplify the interpretation of mass spectra. Moreover, the absence of labels on the side chains would facilitate the quantification of post-translational modifications (PTMs). However, to date all the reported chemical labeling strategies, including dimethyl labeling, iTRAQ (isobaric tags for absolute and relative quantification), and ICAT (isotope-coded affinity tags), result in the modification of side chains.[2–4] One promising method to achieve the incorporation of a single tag is enzymatic labeling. Proteolytic ^{18}O labeling can specifically label tryptic peptide termini without modification of side chains, but the small change in mass and $^{18}O/^{16}O$ back-exchange, i.e. the exchange of ^{18}O with ^{16}O after the labeling process hinder its wide application in quantitative proteomics.[5] Herein, we report a novel enzymatic labeling approach, in which trypsin is used as a ligase to specifically incorporate amino acids labeled with stable isotopes onto the N termini of peptides for quantitative analysis.

Trypsin is a serine protease that specifically hydrolyzes peptide bonds in proteins after arginine and lysine residues. However, trypsin also catalyses peptide synthesis in organic solvents.[6] Therefore, it is possible to covalently link isotopically labeled amino acids to tryptic peptides by using trypsin as a ligase. In this study, arginine, the prototype substrate for trypsin, was used as an acyl moiety donor. The primary amine group of arginine was protected with a benzoyl group (Bz) to prevent the formation of dipeptides or oligopeptides, and the carboxy group was esterified with ethanol to activate the acyl donor (see the Supporting Information, Figure S1). The final product, N_α-benzoyl-L-arginine ethyl ester (Bz–R–OEt), was used as a substrate for the trypsin-catalyzed ligation. The quantitative proteomics workflow based on the trypsin-catalyzed N-terminal labeling is shown in Figure 1a. Trypsin was first used as a protease to digest proteins in an aqueous solution (1M urea/50 mM Tris-HCl, pH 8.0). Then, the generated tryptic peptides were lyophilized and transferred to an ethanol solution containing 4% aqueous buffer (0.1M Tris-HCl, pH 8.0). Next, trypsin immobilized on magnetic nanoparticles (IM-trypsin) and Bz–R–OEt were added to the ethanol solution for the N-terminal labeling of peptides with Bz–R based on a kinetically controlled mechanism (see the Supporting Information). Finally, quantification of the proteins was achieved by the differential labeling of two samples by using Bz–R–OEt (light label) and Bz–($^{13}C_6$)R–OEt (bearing six ^{13}C atoms; heavy label), the incorporation of which are indicated by mass shifts of 260 and 266 Da, respectively.

We first validated the protease and ligase activities of trypsin by using the synthetic peptide VGKANEELAGVVAEVQK (Figure 1b; $m/z = 1740.86$), which contains one trypsin cleavage site. In aqueous solution, treatment with IM-trypsin generated a shorter peptide ANEELAGVVAEVQK (Figure 1b, $m/z = 1456.82$). Treatment of the obtained peptide with an ethanol solution containing Bz–R–OEt and IM-trypsin gave an N-terminus labeled peptide Bz–RANEELAGVVAEVQK (Figure 1b, $m/z = 1716.9$). Similar results were also obtained for three other synthetic peptides that contained one trypsin cleavage site (Supporting Information, Figure S2). These examples clearly illustrate that trypsin functions as a protease in an aqueous solution whereas it acts as an N-terminus ligase in an ethanol solution. Thus, these results imply that this enzymatic labeling approach can be used to label peptides generated by trypsin digestion of a proteome sample. In this study, free trypsin was used for the digestion of proteins, as is conventional in proteomics analysis, whereas IM-trypsin (trypsin from the same source) was used for ligation because it is more tolerant to organic solvents and it is readily removed using magnetism.

We tested whether the labeling of side-chain primary amino groups would also be facilitated by trypsin ligase. A synthetic peptide VIFIEHAKRKG, containing two side-chain amino groups and one terminal primary amino group, was labeled by trypsin. The Arg tag was incorporated only onto the terminal primary amino group and not onto the side-chain amino groups (see the Supporting Information, Figure S3a). These results are markedly different from those of other amine labeling approaches. For example, for labeling

[*] Y. Pan, Prof. Dr. M. Ye, Dr. L. Zhao, K. Cheng, M. Dong, C. Song, H. Qin, Dr. F. Wang, Prof. Dr. H. Zou
CAS Key Lab of Separation Sciences for Analytical Chemistry
National Chromatographic R&A Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences
Dalian 116023 (China)
E-mail: mingliang@dicp.ac.cn
hanfazou@dicp.ac.cn
Y. Pan, K. Cheng, M. Dong, C. Song, H. Qin
University of Chinese Academy of Sciences
Beijing 100049 (China)

[**] This work is funded in part by the Creative Research Group Project of NSFC (21021004), NSFC (21235006, 21275142), the China State Key Basic Research Program Grant (2013CB911202, 2012CB910101, 2012CB910604), and National Key Special Program on Infection diseases (2012ZX10002009-011).

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201303429.

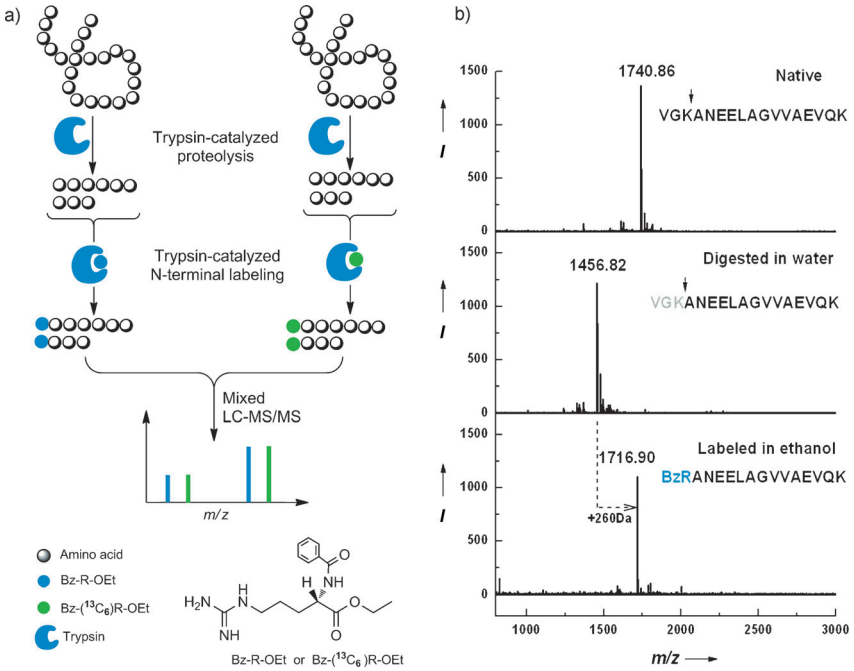

Figure 1. a) Workflow for the trypsin-catalyzed N-terminal labeling strategy. b) Example of the process using a synthetic peptide. 100 μg of the peptide VGKANEELAGVVAEVQK was first digested by IM-trypsin (0.1 mg of IM-trypsin containing 5 μg trypsin) in 50 μL of an aqueous solution (50 mM Tris-HCl, pH 8.0) for 4 h. Then 10 μg of the cleaved peptide was dried and labeled with light Bz-R-OEt (0.16 nmol) using IM-trypsin (0.1 mg containing 5 μg trypsin) in a 50 μL ethanol solution containing 4% aqueous solution of 0.1 M Tris-HCl (pH 8.0) for 6 h. MALDI-TOF MS analysis at each step in the process is provided.

with isotopically labeled dimethyl groups, the three primary amino groups of this peptide, VIFIEHAKRKG, were all dimethylated (see the Supporting Information, Figure S3b).

We were interested to know if the IM-trypsin has residual proteolytic activity in the ethanol solution in which the ligation reaction takes place. The five synthetic peptides used in the initial studies described above were incubated with IM-trypsin in the ethanol solution for 12 h in the absence of Bz–R–OEt. No hydrolysis products were detected after treatment of any of these synthetic peptides (see the Supporting Information, Figure S4), thus indicating that the hydrolysis activity of trypsin under these reaction conditions is negligible. The reaction rate for hydrolysis in the ethanol solution is slow according to chemical kinetics because the concentration of water, which is required for hydrolysis, is low (4%). This may be one of the major reasons for the low hydrolysis activity of trypsin in ethanol.

We also investigated whether minute amounts of sample can be labeled by our approach; as low as 100 fmol of synthetic peptides could be labeled (see the Supporting Information, Figure S5). Finally, this labeling strategy was by using it on myoglobin. Although the digested protein mixture has many peptides containing lysine residues, only one Arg tag was attached per peptide; this result indicates the high specificity of the N-terminal labeling approach (see the Supporting Information).

This labeling strategy was then applied to label a more-complex proteomic sample. Tryptic peptides (100 μg; from mouse liver tissues) generated by free trypsin were labeled with an Arg tag by incubation with Bz–R–OEt (1.6 nmol) and IM-trypsin (1 mg containing 50 μg of trypsin) in 50 μL ethanol solution for 10 h. Then 1 μg of the labeled sample was submitted to LC–MS/MS analysis on LTQ MS instrument. The acquired MS data were searched against the mouse International Protein Index database (IPI-DB) and the Arg tag was set as a variable modification. The new peptide bonds will also be dissociated during fragmentation because arginine was linked to the N termini of the peptides via the formation of peptide bonds. Therefore, the conventional search strategy may lack sensitivity because the extra fragment ions derived from the Arg tag are not considered. To circumvent this problem, we generated a peptide-centric database (PC-DB).[7] All the in silico tryptic peptide sequences with an added N-terminal arginine were included for the identification of ligated peptides. So that this database could be used to identify nonligated peptides, all the in silico tryptic peptide sequences without an added N-terminal arginine were also included (see the Supporting Information). The false discovery rate (FDR) was kept below 1% by using the internal Mascot decoy database search function. About 53% more unique peptides were identified when we searched against the PC-DB compared to the IPI-DB (see the Supporting Information, Figure S6c). Moreover, 97% of the peptides identified using the IPI-DB were also observed using the PC-DB. We manually verified the newly identified peptides, and we found that the use of the PC-DB did not introduce artifacts into the searches. The significant improvement in identification sensitivity by using the PC-DB is attributed mainly to the fact that more MS/MS fragments are considered during the database search (see the Supporting Information, Figure S7). This conclusion is illustrated by the higher Mascot ion scores observed for individual peptides when the PC-DB was used, compared with when the IPI-DB was used (Supporting Information, Table S1). The PC-DB was used to identify labeled peptides in the following studies because of its improved performance.

The label efficiency was determined by calculating the percentage of labeled peptides identified by database searches. Two labeling experiments were performed by labeling the same amount of mouse liver tryptic peptide samples (100 μg each) under the same conditions. Then 1 μg of the obtained samples from the two parallel experiments were analyzed separately by RPLC-MS/MS, thus leading to the identification of 314 and 338 unique peptides, including 285 and 304 labeled peptides, respectively. Therefore, about 90% of the peptides were labeled by Arg tags. To investigate the influence of the Arg tag on the fragmentation of the labeled peptides in collision-induced dissociation (CID), we also searched the MS identification file against the PC-DB for

the nonlabeled mouse liver tryptic peptide sample. Only the N-terminal amino acid residue was different, therefore the majority of y ions from the labeled samples were the same as those of the native peptides (see the Supporting Information, Figure S8). However, there were two marked differences. First, the Arg tag at the N terminus of the labeled peptides dissociated during CID fragmentation as expected. Secondly, the labeled peptide has more continuous b ions, a result that is consistent with response enhancement of b ions after peptide N-terminus labeling by basic arginine.[8] We compared the Mascot ion scores for the 114 peptides identified both in the native sample and in the labeled samples, and found that the majority of the scores were improved after labeling (see the Supporting Information, Table S2). Thus, higher quality MS/MS spectra were obtained after labeling.

To investigate the feasibility of this method for quantitative proteomics, two mixtures of standard proteins were prepared with protein abundance ratios for b-casein, ovalbumin, BSA, myoglobin, and cytochrome c of 1:10, 1:3, 1:1, 4:1, and 7:1, respectively (the quantities of one protein mixture for b-casein, ovalbumin, BSA, myoglobin, and cytochrome c were 1 μg, 1 μg, 1 μg, 4 μg and 7 μg, another were 10 μg, 3 μg, 1 μg, 1 μg and 1 μg). One mixture labeled with the heavy label and the other one with the light one. After combining the labeled samples, the mixture was submitted to LC–MS/MS analysis using an LTQ-Orbitrap MS. All the five proteins were unambiguously identified and accurately quantified with MaxQuant (see the Supporting Information, Table S3). The error differences between the observed and expected quantities ranged between 3% and 10%, thus indicating good correlation of the experimental data with the theoretical design. To explore the applicability of this method to complex proteome samples, one sample of tryptic digests of proteins (100 μg) from mouse liver tissues was labeled with light Arg tags and another sample of the same amount with heavy Arg tags and then they were combined. We analyzed 1 μg of the mixture by LC–MS/MS, and we were able to distinguish that there were 823 unique proteins, achieving an average ratio of 1.04 ± 0.25. Analysis from a replicate labeling experiment revealed 850 unique proteins, with an average ratio of 0.93 ± 0.19. The average ratios (heavy/light) of all proteins for the above two replicate experiments were very close to the theoretical ratio (1:1), thus indicating the high accuracy of this labeling strategy.

Next, this enzymatic isotopic labeling approach was applied to the differential analysis of proteins extracted from hepatocellular carcinoma (HCC) and normal human liver tissues. We were able to reliably quantify 699 unique proteins (Supporting Information, Figure S9d). Many protein ratios quantified in this study were consistent with previous protein quantification datasets obtained using different labeling strategies, such as labeling with isotopically labeled dimethyl groups[9] and iTRAQ[10] (see the Supporting Information, Table S4). Overall, the accuracy of this method was comparable to other chemical labeling methods, and the reliability was better than iTRAQ. Thus, this new labeling strategy is suitable for the analysis of complex proteome samples.

This terminal labeling approach can be combined with de novo peptide sequencing. As shown in Figure 2a, two identical mouse liver tryptic peptide samples (100 μg each) were labeled, one with light Arg and the other with heavy Arg and then were mixed in a 1:1 ratio. The mixture (1 μg) was analyzed by LC–MS/MS using LTQ-Orbitrap with HCD fragmentation. To obtain the simultaneous fragmentation of light and heavy peptides by HCD in the same scan, an isolation width of 8 m/z was set for precursor selection. As the N termini of peptides are isotopically labeled, only N-terminal ions appear as double peaks, thus enabling two ion series to be distinguished. A typical HCD spectrum obtained from one of the labeled peptides is depicted in Figure 2b. Many double peaks with the mass difference of 6.0201 Da were observed. To obtain a simplified spectrum only the light ions of these double peaks were considered, resulting in a spectrum containing only b ions. For such a simplified spectrum, it is straightforward to "read out" the peptide sequence. Moreover, the single peaks in the original spectrum could be collated to form another simplified spectrum containing mainly y ions; this simplified spectrum could also be used for de novo peptide sequencing (Figure 2c). The software for peptide de novo sequencing that uses single-ion series spectra is not available, therefore we searched the two types of spectra constructed from all the acquired HCD spectra against the PC-DB by using Mascot to assess their complementarity for peptide identifications. Only 1153 peptides (29.2%) were identified in both the b- and y-ion spectra (Figure 2d), thus indicating that these two types of spectra are highly complementary. Owing to the presence of highly basic groups at the C terminus, the fragmentation of native tryptic peptides preferentially produces y ions. However, for the Arg-labeled tryptic peptides more b ions were generated. A similar number of labeled peptides (1319 compared to 1481, Figure 2d) were identified by the b- and y-ion spectra, thus indicating that the two types of ions are equally useful for peptide identification. Terminal modification[11] or digestion with a special protease[12] can directly generate spectra with a single ion series. However, such methods only generate one type of ion series, whereas the N-terminal isotopic labeling method is able to generate two types of single ion series spectra, which could be used together for de novo sequences (see the Supporting Information, Figure S10). This is a distinct advantage over other methods because the single ion series spectrum often has gaps.

In summary, this study demonstrates a unique approach for quantitative proteomics based on the specific N-terminal labeling of tryptic peptides under mild conditions. Although we focused on isotopically labeled amino acids, different isotopically labelled Bz–R–OEt could be easily developed by using either isotopically labeled amino acids or blocking groups. In addition to trypsin, a variety of proteases with different specificities have been reported to have peptide synthesis activity under appropriate conditions.[13] Therefore, specific ligation of other amino acids to the N or C termini of tryptic peptides for quantification of proteins might also be feasible. The diversity of possible isotopic tags and proteases

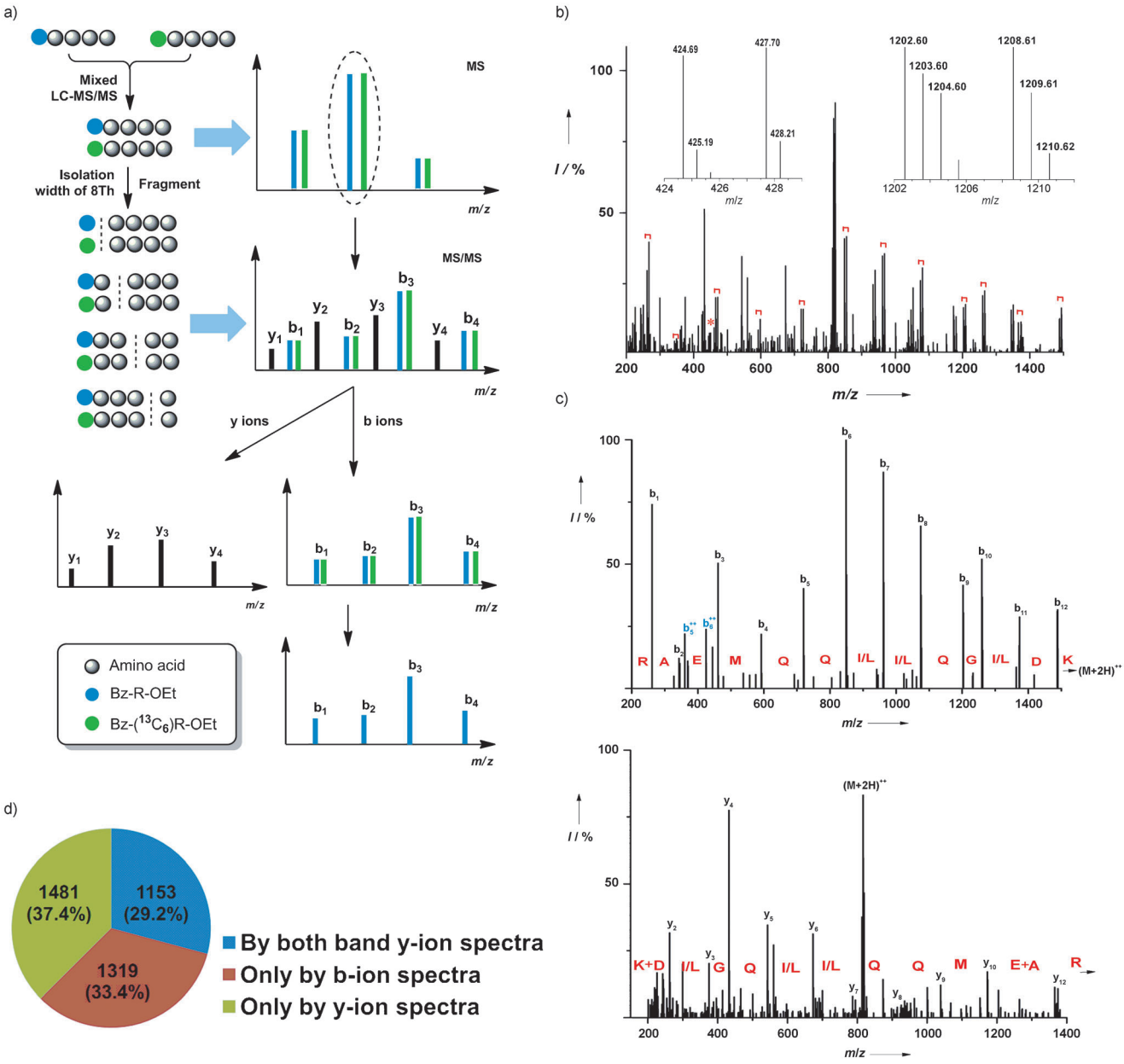

Figure 2. Determination of ion series for de novo sequencing by differential N-terminal isotopic labeling. a) Schematics for the determination of ion series. The b ions appear as double peaks, whereas the y ions appear as single peaks. An MS/MS can be deconstructed into two single ion series spectra, that is, a b ion spectrum and a y ion spectrum. b) The HCD spectrum for a representative isotopically labeled peptide from mouse liver digest. All ions that do not contain the N-terminal label are single peaks, whereas the b ions in the HCD spectrum appear as double peaks, as indicated. c) The b- and y-ion spectra regenerated from the HCD spectrum. d) The comparison of peptide identifications achieved by searching the b-ion spectrum and y-ion spectrum against the database with Mascot.

with different ligation specificity makes this labeling approach a promising strategy for quantitative proteomics.

Received: April 23, 2013
Published online: July 5, 2013

Keywords: enzyme catalysis · isotopic labeling · peptides · quantitative proteomics · terminal labeling

[1] S. E. Ong, M. Mann, *Nat. Chem. Biol.* **2005**, *1*, 252–262.
[2] J. L. Hsu, S. Y. Huang, N. H. Chow, S. H. Chen, *Anal. Chem.* **2003**, *75*, 6843–6852.
[3] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, D. J. Pappin, *Mol. Cell. Proteomics* **2004**, *3*, 1154–1169.
[4] S. P. Gygi, B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb, R. Aebersold, *Nat. Biotechnol.* **1999**, *17*, 994–999.

[5] a) O. A. Mirgorodskaya, Y. P. Kozmin, M. I. Titov, R. Körner, C. P. Sönksen, P. Roepstorff, *Rapid Commun. Mass Spectrom.* **2000**, *14*, 1226 – 1232; b) X. Yao, A. Freas, J. Ramirez, P. A. Demirev, C. Fenselau, *Anal. Chem.* **2001**, *73*, 2836 – 2842; c) B. O. Petritis, W.-J. Qian, D. G. Camp, R. D. Smith, *J. Proteome Res.* **2009**, *8*, 2157 – 2163.
[6] a) T. Oka, K. Morihara, *J. Biochem.* **1977**, *82*, 1055 – 1062; b) F. Bordusa, *Chem. Rev.* **2002**, *102*, 4817 – 4868.
[7] C. Y. Yen, S. Russell, A. M. Mendoza, K. Meyer-Arendt, S. J. Sun, K. J. Cios, N. G. Ahn, K. A. Resing, *Anal. Chem.* **2006**, *78*, 1071 – 1084.
[8] E. Ernoult, E. Gamelin, C. Guette, *Proteome Sci.* **2008**, 6:27.
[9] F. Wang, R. Chen, J. Zhu, D. Sun, C. Song, Y. Wu, M. Ye, L. Wang, H. Zou, *Anal. Chem.* **2010**, *82*, 3007 – 3015.
[10] R. Chaerkady, H. C. Harsha, A. Nalli, M. Gucek, P. Vivekanandan, J. Akhtar, R. N. Cole, J. Simmers, R. D. Schulick, S. Singh, M. Torbenson, A. Pandey, P. J. Thuluvath, *J. Proteome Res.* **2008**, *7*, 4289 – 4298.
[11] T. Keough, R. S. Youngquist, M. P. Lacey, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 7131 – 7136.
[12] N. Taouatas, M. M. Drugan, A. J. R. Heck, S. Mohammed, *Nat. Methods* **2008**, *5*, 405 – 407.
[13] a) K. M. Koeller, C. H. Wong, *Nature* **2001**, *409*, 232 – 240; b) C. Lombard, J. Saulnier, J. Wallach, *Protein Pept. Lett.* **2005**, *12*, 621 – 629.

